SYSTEM AND METHOD FOR EVALUATING DATA SETS OVER A COMMUNICATIONS NETWORK

BACKGROUND OF THE INVENTION

1. Field of Invention

The present invention relates to computer-implemented data management and analysis systems and methods. In particular, the present invention is related to a system and method for warehousing, managing, and analyzing large time-series and non time-series data files stored on a server by collaborative researchers who are located at remote locations but who are in data communication with the server.

2. Description of the Related Art

Collaboration among research scientists using distributed client computers over a communication network is not new. For example, U.S. Pat. No. 6,611,822 describes a collaborative session that involves establishing a network connection between a plurality of users, selecting a mode for the network connection between the plurality of users, establishing a network connection mode between the plurality of users, and synchronizing the mode between the plurality of users. The patent teaches that the modes of operation include application and information sharing over the network.

Often, collaborative research involves analyzing data sets. U.S. Pat. No. 6,615,253 describes a system involving server side data retrieval for execution of client side applications. The patent teaches a method of requesting data stored on a server over a network, bundling the data into a data structure in response to the request, and sending the data structure to a client computer over the network, whereby the data structure is cached on the client and used as needed during execution of the application running on the client.

Client applications for analyzing data have been around for may years. One of the more robust applications, MATLAB®, provides a "distributed computing toolbox" and a "distributed computing engine" that enable users to develop distributed analysis applications and execute them over a cluster of different computers (presumably even ones geographically remote from each other) without leaving a central development environment where the user is located. The data set, or a portion thereof, must still reside on each of the distributed computers, and the analysis module inputs for data mining are created from scratch by the user.

Similarly, simple spreadsheet programs like Microsoft® Excel® can be used to store, in a file on a client computer, a set of data records that can be shared by multiple users over a network so that collaborators can each manipulate the data for his or her use. However, this requires each collaborator to develop data mining queries, conduct mathematical operations, and analyze the results on his or her own computer.

While the aforementioned data analysis and collaboration techniques, and others like them, are feasible for relatively small data sets, the method is not suitable for very large data sets, especially time-series data sets that can easily range between tens to hundreds of gigabytes of data and require the collaboration of domain experts in diverse fields engaged in the research process. Traditionally, as taught by the above patents, the data sets need to be electronically distributed to the collaborators so they can analyze them using analytical applications running on their own computers. Thus, each research team had to locally maintain the tools necessary to properly handle and store the data. Moreover, facilities for querying and analyzing the data would need to have been developed by each researcher using the application on his or her client compute. Furthermore, establishing common facilities among the various collaborative groups and overcoming issues of hardware and software incompatibility would need to have been addressed. Also, the lack of an integrated system for managing and analyzing large sets of time-series data forced collaborators to develop endless data input-output interface systems for sequential data analysis.

U.S. Pat. No. 6,405,195 addresses some of those problems. It describes a system in which data to be analyzed is transferred from one or more user systems to a host system, which includes an analysis/decision support module. Queries are generated, either automatically by the analysis/decision support module, or by the user, who then submits them to the host system. As taught in the patent, more than one user may participate in the system, including transferring data to the host. This joint participation includes the option of collaboratively submitting or adjusting queries and viewing the results of the data analysis, either in real time, or asynchronously. Data used as the basis of an analysis may therefore come from different entities, even from databases that are available publicly via the network, but whose owners are not participants in the collaborative, hosted analysis system according to the disclosed invention. The patent also describes how the host system acts as a network portal through which different users may store and share not only data for analysis, but also the results of such analysis. Notwithstanding the above, the patent does not deal with very large time-series data records, which present unique challenges over conventional database management systems.

Existing database management systems are not optimal for the management of very large time series data, since they were not built with this objective. Their storage architecture is not structured to effectively handle ordered data and simple time series operations are poorly supported. Querying the data requires writing statements in some form of query language; this is acceptable in business applications where the same queries are performed over and over again, but becomes burdensome when the database must be subjected to ad hoc queries. The burden, then, on collaborative researchers has traditionally been placed on the research team itself to construct a solution from scratch, resulting in each research study requiring its own effort to combine data storage, search, and analysis in order to reach certain research objectives.

Accordingly, there exists the need for a system and method for evaluating data over a communications network that is an efficient and effective collaboration tool for remote researchers. The present invention allows researchers to transition away from the mundane yet necessary task of data and system management that has plagued the prior art systems and methods of time-series data evaluation, and to focus on their core research objectives: data mining and analysis.

SUMMARY AND OBJECTS OF THE INVENTION

The Combat Casualty Care Directorate of the U.S. Army Medical Research and Materiel Command (USAMRMC), Ft. Detrick, Md., supports various intramural and extramural research studies involving the collection and analysis of human and animal physiological data. Those studies generate voluminous amounts of time-series data, such as, but not limited to, electrocardiogram waveforms (EKG), oxygen saturation waveform (SpO2), and respiratory traces, that traditional relational database management systems are ill-suited to handle. Therefore, those data sets are generally stored by investigators in their individual workstations and are manually manipulated for subsequent visualization and analysis.